Eukaryotic start and stop translation sites

Douglas R.Cavener* and Stuart C.Ray+
Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA

Received December 18, 1990; Revised and Accepted May 15, 1991

ABSTRACT

Sequences flanking translational initiation and termination sites have been compiled and statistically analyzed for various eukaryotic taxonomic groups. A few key similarities between taxonomic groups support conserved mechanisms of initiation and termination. However, a high degree of sequence variation at these sites within and between various eukaryotic groups suggest that translation may be modulated for many mRNAs. Multipositional analysis of di-, tri-, and quadri-nucleotide sequences flanking start/stop sites indicate significant biases. In particular, strong tri-nucleotide biases are observed at the −3, −2, and −1 positions upstream of the start codon. These biases and the interspecific variation in nucleotide preferences at these three positions have lead us to propose a revised model of the interaction of the 18S ribosomal RNA with the mRNA at the site of translation initiation. Unusually strong biases against the CG dinucleotide immediately downstream of termination codons suggest that they may lead to faulty termination and/or failure of the ribosome to disassociate from the mRNA.

INTRODUCTION

The messenger RNA database has increased by more than an order of magnitude since Kozak (1, 2) first reported CCA/GCC-AUGG as a consensus sequence for the eukaryotic translation initiation site. This consensus sequence was derived largely from vertebrate mRNAs. More recent compilations of start codon flanking sequence for Drosophila (3), yeast (4, 5), and plants (6) have shown that a considerable degree of inter-taxon variation exists. However, all eukaryotic groups examined to date share a strong preference for purines at the −3 position upstream of the start codon suggesting a conserved translation initiation mechanism. Genetic experiments indicate that the −3 position has a variable degree of importance to translation depending upon the species examined (7−12) and the sequence of other positions, particularly at −2, −1 and +4 (7, 12). mRNAs translated in mammalian and Drosophila systems appear to be more sensitive to mutations which introduce non-consensus sequences (7, 8, 12) than yeast.

In contrast to translation initiation, translation termination in eukaryotes has received less attention. Valle and Morch have recently reviewed eukaryotic termination (13) and argued that regulation of protein synthesis at the level of termination may be more common than once thought. In particular, failure to read termination codons as such can result in extending the carboxyl terminus of the nascent protein or to shift to another reading frame. These seemingly unusual events are necessary for the synthesis of essential proteins for many viruses (14, 15). In addition, under certain circumstances ribosomes have the ability to terminate and reinitiate translation, apparently without disassociating from the mRNA (16−20). The circumstances leading to termination-disassociation versus termination-reinitiation have not been investigated. A compilation mRNA eukaryotic termination sequences suggested a preference for purines immediately flanking the 3′ side of stop codons (21), in contrast to prokaryotes which general have a strong preference for U at this position (22).

We present a comprehensive analysis of sequences flanking the start and stop codons for vertebrates, vertebrate viruses, yeast, Drosophila, invertebrates (not including Drosophila), protozoans, Dictyostelium, monocot plants, and dicot plants. Since the ribosome interacts simultaneously with several nucleotides in an mRNA, synergistic effects among nucleotide surrounding the start codons is a distinct possibility. Heretofore, the eukaryotic mRNA database has been too small to perform a multiposition-statistical analysis which might provide circumstantial evidence for such synergism. The eukaryote database is now large enough to statistically examine doublet, triplet and quadruplet nucleotide combinations. We have designed a computer program named Interbas to compile and statistically analyze multiple positions while factoring out the normal doublet, triplet, and quadruplet biases which occur in mRNA sequences. A plethora of statistical interactions between nucleotides and sequence positions have been discovered. Most of these interactions can be explained by differences in GC content bias between mRNA, but some are likely the result of natural selection for modulating initiation or termination of translation. The discovery of strong triplet biases immediately upstream of start codons has lead us to propose a modified model for the interaction of mRNA and the 18S rRNA.

* To whom correspondence should be addressed

+ Present address: Tower 110, Johns Hopkins Hospital, Baltimore, MD 21205, USA

METHODS

Interbas program and compilation of sequences

A computer program dubbed Interbas was written to (a) compile sequences flanking translational termination and initiation sites from the Genbank database (Release 63), (b) edit sequence files and delete redundancies and closely related homologs, (c) calculate marginal nucleotide frequencies, and (d) statistically analyze multiple nucleotide positions. Interbas is a menu-driven program written in Turbo Pascal (Borland International, Inc.) for use with Genbank on MS/PC-DOS computers. Interbas was used to compile Genbank sequences from −23 to +6 flanking translation initiation codons and sequences from −9 to +13 flanking translation termination codons. Since the statistical analyses performed assumes that each sequence is an independent random variable, considerable editing of the databases was required. Redundant mRNA sequences were automatically flagged and eliminated. In addition the compiled sequences were manually scanned at least twice to identify and eliminate mRNA sequences which share a close homologous (paralogous or orthologous) relationship with another mRNA sequence in the database. For example, a large number of mRNA sequences encoding closely related MHC antigens and globins were eliminated. Only human sequences were retained in the primate database since virtually all of the closely related non-human primate sequences in Genbank were isolated on the basis of their homology with human genes (which also exist in the database). For the same reason, only Drosophila melanogaster sequences were used for the Drosophila database. The vertebrate virus database required extensive editing because it contained a very large number of closely related genes from various virus isolates.

Multipositional statistical analysis

Expected frequencies of doublets, triplets, and quadruplets were calculated on the basis of independence of each sequential nucleotide position relative to the start or stop codon corrected for doublet, triplet, and quadruplet frequency biases which generally occur in the untranslated region of the sampled mRNAs. Expected doublet, triplet, and quadruplet frequencies were calculated as the products of $X_iX_jC_D$, $X_iX_jX_kC_T$, and $X_iX_jX_kX_lC_Q$, respectively, where X_n = the frequency of A, G, C, or U at the i, j,k, or l position relative to the stop or start codon and C_D, C_T, and C_Q = the doublet, triplet, and quadruplet correction factors, respectively. The correction factors C_D, C_T, and C_Q were calculated as the ratio of the observed doublets, triplets, and quadruplets to the expected doublets, triplets, and quadruplets summed and averaged for more than eight positions. For the translation initiation site the correction factors were calculated from the region −20 to −10 and for the termination site the region +5 to +13 was used.

Non-sequential doublets, triplets, and quadruplets were also analyzed. Doublet and triplet correction factors were used for non-sequential multipositional analysis only when part of a triplet or quadruplet contained a sequential doublet or triplet (e.g. $X_i.X_kX_l$).

50/75 Consensus Rule

The criteria for consensus sequence assignment is after Cavener (3) and will be referred to as the 50/75 Consensus Rule. If a nucleotide at specific position occurs in more than 50% of the sequences and is greater than twice the frequency of the next most frequent nucleotide, it is assigned as the sole consensus nucleotide. If the sum of the frequencies of the two most frequent nucleotides is greater than 75% (but neither meet the criteria as a sole consensus nucleotide) they are assigned as co-consensus nucleotides. If no single nucleotide or pair of nucleotides meet these consensus criteria the letter N is assigned to that position to indicate lack of consensus or a lower case letter is used to indicate the most frequent nucleotide present.

RESULTS

Nucleotide frequencies of the translation initiation site

The translation initiation sequences contained in the primate, rodent, other mammalian, and other vertebrate sections of Genbank Rel 63 were compiled separately. Since these four groups yielded very similar nucleotide frequency distributions, they were consolidated as a single all vertebrate data base containing 2,595 mRNAs (Table 1). Curiously, as the vertebrate data base has risen from 179 sequences (2,3) to 699 (23) and now to 2,595, the frequency of A at the −3 position has dropped from 78% to 61% and now to 58%. A smaller degree of erosion has occurred at the other position including the −4 position where the frequency of C is now below 50%. The consensus sequence for the vertebrate translation initiation site is A/GNCAUG as

Table 1 Nucleotide frequencies flanking eukaryotic start codons

		-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5
Vertebrates	A	25	22	19	23	20	19	25	58	28	17		25	24
n = 2,595	G	24	32	24	22	42	24	18	33	16	23		46	22
	C	31	27	38	33	20	37	48	6	45	53		15	37
	U	21	19	19	21	18	20	9	3	12	7		13	17
		c	g	c	c	g	c	c	A/G	c	C	AUG	g	c
Vertebrate	A	30	29	26	27	26	25	29	56	40	32		23	37
viruses	G	21	25	21	20	27	17	17	28	11	22		53	17
n = 349	C	27	19	30	23	20	30	38	8	29	33		9	34
	U	23	27	24	30	27	28	16	9	20	13		14	11
		a	a	c	u	g/u	c	c	A/G	a	c	AUG	G	a
Drosophila	A	31	38	26	34	32	22	21	65	47	39		26	28
n = 192	G	19	20	16	16	28	19	12	20	12	19		35	21
	C	34	17	34	29	17	36	53	7	23	34		19	30
	U	17	25	25	21	23	22	14	8	18	8		20	20
		c	a	c	a	a	c	C	A	a	a	AUG	g	c
Other	A	45	30	30	46	37	23	41	77	46	34		34	34
invertebrates	G	17	17	15	12	24	15	9	16	7	12		34	19
n = 155	C	16	21	28	21	13	25	38	2	20	40		15	38
	U	23	31	26	21	26	37	12	5	26	14		18	9
		a	u	a	a	a	u	A/C	A	a	c	AUG	a/g	c
Monocot plants	A	35	23	29	33	33	26	37	42	23	23		15	15
n = 93	G	24	31	18	33	39	22	29	46	18	31		71	12
	C	25	24	30	25	15	38	30	9	51	44		1	70
	U	16	23	23	9	13	15	4	3	9	2		13	3
		a	g	c	a/g	g	c	a	A/G	C	c	AUG	G	C
Dicot plants	A	35	35	37	41	40	35	49	70	42	50		14	16
n = 233	G	21	13	13	14	21	12	16	17	5	11		70	12
	C	16	21	24	16	14	27	22	6	32	27		3	63
	U	28	32	26	29	25	25	13	8	21	12		13	9
		a	a	a	a	a	a	a	A	a	A/C	AUG	G	C
Yeast	A	41	40	46	42	35	35	43	66	46	43		26	23
n = 461	G	10	14	11	13	18	12	10	18	9	15		28	12
	C	17	17	21	17	17	22	26	7	22	20		12	43
	U	31	29	22	28	30	30	21	8	23	21		34	22
		a	a	a	a	a	a	a	A	a	a	AUG	u	c
Dictyostelium	A	58	60	60	56	67	69	67	93	84	91		44	42
n = 45	G	4	4	9	9	2	0	0	2	0	4		33	22
	C	7	9	7	7	2	11	9	0	7	2		7	24
	U	31	27	24	29	29	20	24	4	9	2		16	11
		A	A	A	A/U	A	A	A	A	A	A	AUG	A/G	a
Protozoa	A	34	32	32	38	31	34	47	71	50	51		40	30
n = 112	G	10	4	18	11	23	10	15	12	11	18		27	19
	C	20	18	16	17	13	23	20	8	23	18		12	32
	U	37	46	34	34	32	33	18	10	16	13		21	19
		u	u	u	a	u	a	a	A	A	A	AUG	a	c

determined by the application of the 50/75 consensus rule. This consensus sequence is found in 1,277 (49%) of the vertebrate mRNAs. In all other positions from −10 to −1 upstream of the start codon, C is preferred except at −9 and −6 where G is preferred. At positions +4 and +5, and G and C are preferred, respectively. Although the major classes of vertebrates are generally homogeneous with respect to nucleotide frequencies flanking start sites, the viruses isolated from vertebrates exhibit a distinct frequency profile. Consequently the vertebrate virus mRNA sequences were compiled separately (Table 1). Vertebrate virus mRNAs are significantly more AU biased than the vertebrate cellular mRNAs.

A comparison of Drosophila start sequences with those of other invertebrates indicated a significant differences at the −4 position. Consequently these two groups were separately compiled (Table 1). Similarly, monocot plants and dicot plants differ substantially at a number of positions and were therefore separately compiled. Not unexpectedly, the most pronounced feature of all of the eight eukaryotic groups compiled is the strong preference for A or G at the −3 position. The combined A + G percent at this position ranges from 95% (in Dictyostelium) to 83% in protozoa. Although the frequency of A at the −3 position is typically 2 to 4 fold higher than G, monocot plant actually exhibit a higher frequency G than A. Monocots exhibit an unusually high G frequency throughout the −10 to −1 region. This is in contrast to Dictyostelium, which is G and C depauperate in this region. In addition to the −3 position, the −4, −2, and −1 positions upstream of the start codon typically exhibit consensus nucleotides. When a consensus exists at these three positions, it is either for A or C.

Mutagenesis experiments have indicated that the substitution of T at positions −3, −2, and −1 depresses translation in mammalian and Drosophila translation systems (7,12). The frequency of U is generally low at the −3 and −1 position, (Table 1) but in several of the eukaryotic groups an appreciable frequency of U occurs at −2. The relative frequency of U in Dictyostelium is particularly interesting because the overall AU bias is abruptly switched to an A-only bias beginning at the −3 position. Yeast are exceptional in exhibiting relative high U frequencies at both the −2 and −1 positions.

Kozak (23) noted a periodic increase in the frequency of G at position −9, −6, and −3 in vertebrates. This periodicity is a generally observed at least for the −6 and −3 positions for all of the eukaryotic groups with the exception of Dictyostelium (Table 1). Downstream of the start codon, both monocot and dicot plants exhibit a strong preference for alanine codons (GCN).

Since the −3 position appears to be the most crucial position upstream of start codons, the data sets were subdivided into the four groups based upon the specific nucleotide at the −3 position (Table 2). For this and other multipositional analyses, all of the non-vertebrate taxa (excluding Dictyostelium) were pooled. Vertebrate viruses were not included since their frequency profile is significantly different from cellular vertebrate mRNAs.

In general the frequency profiles of each of the individual −3 subgroups exhibit the characteristic biases detailed above for the unsubdivided data. At the −2, −1 and +4 positions the most frequent nucleotide is identical for all four groups within each of the two taxa (CC...G for vertebrates and AA...G for non-vertebrates). All eight datasets in Table 1 exhibit a relatively high frequency of G at −6 compared with −7, −5 and −4. Inasmuch as Kozak (7) has shown that C or U at −3 substantially reduces the efficiency of translation initiation and C and U are relatively rare at these positions, potential start sites bearing C or U at −3 have often been viewed suspiciously. However, the overall similarity in bias among all four −3 subgroups (Table 2) suggests that pyrimidines at the −3 position may be compatible with initiation of translation.

Multipositional analyses of translation initiation sequences

Triplet frequencies at the −3, −2, and −1 positions (−321) for vertebrate and non-vertebrates were compiled and compared to their expected frequencies (Table 3). Two alternative methods were used to calculate expected frequencies. Method-I expected frequencies were calculated as the product of frequencies over the three positions. For example the expected number of the AAA triplet in vertebrates using this method is (0.577)(0.279)(0.171)(2595) = 71.4. Method-I does not correct for the normal triplet bias which exists in the leader sequence nor does it correct for the bias associated with being adjacent to the invariant A nucleotide at +1. Method-II corrects for these two biases by multiplying the expected number calculated by Method-I by the appropriate quadruple factor derived from an analysis of the leader sequence upstream of the −321 site (−20 to −9). (See materials and methods for a description of this analysis.) Quadruple factors which included an A nucleotide at the fourth position were used instead of triple factors. For example the expected number of the vertebrate AAA triplet using method-II is (71.4)(2.04) = 145.7. Therefore, in this example, the observed frequency of AAA (137) is close to the expected frequency (145.7) relative to the normal AAA triplet bias in the leader sequence. The predicted values from method-II are used for comparative purposes below.

The most frequent −321 triplets are ACC and GCC, for vertebrates, and AAA and ACA for non-vertebrates. For both taxonomic groups AAC is predicted to be the second most frequent triplet, however the observed values are approximately half the expected values. Most of the other NAC triplets also occur at lower than expected values (e.g. vertebrate GAC). Several other −321 triplet frequencies deviate dramatically from

Table 2 Frequency of start sequences grouped according to -3 subgroups

	Vertebrate								Other eukaryotes							
	-7	-6	-5	-4	-3	-2	-1	+4	-7	-6	-5	-4	-3	-2	-1	+4
A	28	21	20	20	100	32	18	29	45	37	33	40	100	48	44	26
G	19	43	25	17	0	19	25	41	14	22	13	14	0	9	15	41
C	31	17	34	57	0	38	51	15	17	15	26	33	0	23	28	11
U	22	19	21	6	0	11	6	14	24	26	27	13	0	20	14	23
	c	g	c	C	A	c	C	g	a	a	a	a	A	a	a	g
A	18	18	14	37	-	22	13	19	29	34	25	44	-	39	36	22
G	26	42	26	19	100	10	19	53	19	27	17	12	100	8	13	48
C	37	23	44	32	-	58	59	16	27	15	29	31	-	35	33	7
U	20	17	16	13	-	11	8	13	25	25	30	13	-	18	18	23
	c	g	c	a	G	C	C	G	a	a	u	a	G	a	a	g
A	17	23	27	21	-	32	23	21	31	31	28	37	-	37	42	25
G	31	30	11	21	-	10	31	62	12	17	12	15	-	10	27	38
C	31	25	31	48	100	44	38	11	26	17	28	16	100	27	21	14
U	21	23	31	9	-	14	8	6	31	35	31	32	-	26	10	23
	g/c	g	c/u	c	C	c	c	G	a/u	u	u	a	C	a	a	g
A	12	28	33	21	-	17	22	26	34	30	24	34	-	30	40	28
G	31	32	19	19	-	27	21	49	14	24	14	12	-	14	26	29
C	31	23	28	42	-	38	47	14	18	20	29	23	-	29	20	17
U	26	16	20	19	100	17	10	11	34	26	33	31	100	28	16	27
	g/c	g	a	c	U	c	c	G/A	a/u	a	u	a	U	a	a	g

Sample size, vertebrates A= 1497, G = 860, C = 154, U = 81;
other eukaryotes A = 829, G = 240, C = 81, U = 94;

Table 3 Triplet frequencies for position -3, -2, -1 upstream of vertebrate and non-vertebrate start sites

	Vertebrates				Non-vertebrates					Vertebrates				Non-vertebrates			
	Obs	%	E-I	E-II	Obs	%	E-I	E-II		Obs	%	E-I	E-II	Obs	%	E-I	E-II
AAA	137	5.3	71	146	**185**	**14.9**	152	254	AAG	203	7.8	96	210	85	6.8	58	81
AAC	114	4.4	222	294	72	5.8	101	139	AAU	19	0.7	29	27	54	4.3	53	48
AGA	49	1.9	40	102	26	2.1	33	49	AGG	126	4.9	54	118	19	1.5	13	16
AGC	92	3.5	125	268	19	1.5	22	31	AGU	19	0.7	16	11	12	1.0	11	11
ACA	65	2.5	115	110	90	7.2	91	110	ACG	31	1.2	154	69	18	1.4	35	20
ACC	**436**	**16.8**	357	351	62	5.0	60	61	ACU	36	1.4	47	28	22	1.8	32	22
AUA	22	0.8	30	24	61	4.9	71	54	AUG	9	0.3	40	11	1	0.1	27	5
AUC	126	4.9	92	70	77	6.2	47	45	AUU	14	0.5	12	8	27	2.2	25	17
GAA	47	1.8	41	81	37	3.0	44	63	GAG	52	2.0	55	113	16	1.3	17	32
GAC	80	3.1	127	159	19	1.5	29	15	GAU	9	0.3	17	8	21	1.7	15	12
GGA	16	0.6	23	40	5	0.4	9	14	GGG	36	1.4	31	39	2	0.2	4	3
GGC	28	1.1	72	67	7	0.6	6	7	GGU	4	0.2	9	4	6	0.5	3	2
GCA	45	1.7	66	59	30	2.4	26	32	GCG	54	2.1	88	35	9	0.7	10	7
GCC	347	13.4	205	220	36	2.9	17	21	GCU	51	2.0	27	15	9	0.7	9	6
GUA	8	0.3	17	9	14	1.1	21	14	GUG	19	0.7	23	22	5	0.4	8	7
GUC	56	2.2	53	39	17	1.4	14	12	GUU	8	0.3	7	2	7	0.6	7	6
CAA	10	0.4	7	7	15	1.2	15	18	CAG	28	1.1	10	23	7	0.6	6	7
CAC	9	0.3	23	27	3	0.2	10	13	CAU	2	0.1	3	1	5	0.4	5	4
CGA	1	<0.1	4	1	5	0.4	3	3	CGG	7	0.3	6	4	0	0.0	1	1
CGC	4	0.2	13	7	2	0.2	2	2	CGU	4	0.2	2	<1	1	0.1	1	<1
CCA	22	0.8	12	12	7	0.6	9	10	CCG	6	0.2	16	9	11	0.9	3	2
CCC	34	1.3	37	40	4	0.3	6	4	CCU	5	0.2	5	4	0	0.0	3	1
CUA	3	0.1	3	2	7	0.6	7	5	CUG	6	0.2	4	8	4	0.3	3	3
CUC	12	0.5	9	10	8	0.6	5	4	CUU	1	<0.1	1	<1	2	0.2	2	2
UAA	4	0.2	4	5	9	0.7	17	14	UAG	5	0.2	5	4	9	0.7	7	4
UAC	5	0.2	12	7	5	0.4	11	12	UAU	0	0	2	<1	5	0.4	6	5
UGA	5	0.2	2	4	8	0.6	4	3	UGG	9	0.3	3	4	3	0.2	1	2
UGC	6	0.2	7	9	1	0.1	3	2	UGU	2	0.1	1	<1	1	0.1	1	<1
UCA	5	0.2	6	5	12	1.0	10	13	UCG	0	0	9	2	7	0.6	4	5
UCC	22	0.8	20	33	5	0.4	7	7	UCU	5	0.2	3	3	3	0.2	4	3
UUA	4	0.2	2	2	9	0.7	8	6	UUG	3	0.1	2	3	3	0.2	3	4
UUC	6	0.2	5	7	8	0.6	5	8	UUU	2	0.1	<1	<1	6	0.5	3	3

Obs = observed number

E-I (Method I): χ^2 = 808.4, P < 0.001, vertebrates; χ^2 = 147.5, P < 0.01, non-vertebrates

E-II (Method II): χ^2 = 725.8, P < 0.001, vertebrates; χ^2 = 198.9, P < 0.01, non-vertebrates

Observed number and percent of the most frequent triplets for vertebrates and non-vertebrates are in bold type.

the expected values. In particular, the observed frequency of all eight of the vertebrate RYY (R = purine and Y = pyrimidine) −321 triplets are greater than expected, whereas the frequency of RRY triplets are generally much less frequent than expected.

The most frequent −321 triplet in vertebrate viruses is AAA; the ACC triplet is the sixth most frequent. This is precisely the same order of these two triplets in the non-vertebrate dataset. However, in the cellular vertebrate dataset ACC is the most frequent −321 triplet whereas AAA is the fifth most frequent triplet. This seeming paradox is most likely the result of the relative AU bias of vertebrate virus mRNAs sampled compared to vertebrate cellular mRNAs.

Kozak (7) has shown that the −3 and +4 positions surrounding the start codon of the preproinsulin gene synergistically effect initiation of translation. In particular the repressive effect of U at −3 could be partially alleviated by the presence of G at +4, whereas G at +4 appeared to have much more minor influence when the −3 position was occupied by other nucleotides besides U. Although the observed frequency of the U...U doublet is low for both groups (Table 4), its frequency is very close to the expected frequency. In general, large statistical interactions between the −3 and +4 nucleotides do not occur as might be expected from a functional synergism.

In addition to the multipositional analysis detailed above, all the possible combinations of doublet, triplet, and quadruplet frequencies throughout the −20 to +6 region surrounding eukaryotic start codons were determined and compared to expected frequencies (data not shown). This analysis involved several thousand comparisons and, not unexpectedly, several observed frequencies deviated significantly from expected frequencies. Most deviations could be accounted for by GC content heterogeneity of the data sets (i.e. combinations of G and C and combinations of A and U were higher than expected). Although some of the observed deviations could be functionally significant, none suggested the presence of a frequent motif analogous to the Shine-Dalgarno sequences found at variable positions upstream of bacterial mRNAs (24). Finally, it should be noted that the strong biases observed at −321 (detailed above) were several fold higher than any other triplet combination in the region from −20 to +6.

Nucleotide frequencies of the translation termination site

The relative usage of the three termination codons (UAA, UAG, and UGA) varies considerably between eukaryotic groups (Table 5). UGA is the most frequent of the three termination codons for vertebrates and monocot plants; whereas UAA is

Table 4. Frequency of -3 and +4 doublets surrounding eukaryotic start codons

	Vertebrates			Non-vertebrates		
-3 +4	Obs	%	Exp #	Obs	%	Exp #
A....A	441	17.0	379.7	214	17.2	208.7
A....G	615	23.7	692.3	336	27.0	338.7
A....C	230	8.9	226.2	91	7.3	90.0
A....U	211	8.1	195.6	189	15.2	192.7
G....A	164	6.3	218.2	53	4.3	60.3
G....G	454	17.5	399.5	114	9.2	97.9
G....C	134	5.2	130.0	17	1.4	26.0
G....U	108	4.2	112.4	56	4.5	55.7
C....A	32	1.2	39.1	20	1.6	20.4
C....G	96	3.7	71.5	31	2.5	33.1
C....C	17	0.7	23.3	11	0.9	8.8
C....U	9	0.3	20.1	19	1.5	18.8
U....A	21	0.8	21.1	26	2.1	23.6
U....G	40	1.5	38.6	27	2.2	38.4
U....C	11	0.4	12.5	16	1.3	10.2
U....U	11	0.4	10.9	25	2.0	21.8

Vertebrates $\chi^2 = 58.7$, $P < 0.01$

Non-vertebrates $\chi^2 = 27.57$, $P < 0.05$

Table 5 Eukaryotic stop codon frequencies

	UAA	UAG	UGA
Vertebrates	33	22	45
Vertebrate viruses	50	24	26
Drosophila	51	28	22
Other invertebrates	60	17	23
Monocot plants	23	27	50
Dicot plants	48	21	31
Yeast	54	17	29
Dictyostelium	91	4	6
Protozoa	51	21	27

Table 6 Nucleotide frequencies flanking eukaryotic stop codons

		-6	-5	-4		+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
Vertebrates	A	28	36	16		34	23	26	19	24	22	24	21	23	22
n= 2,006	G	26	16	27		35	28	27	26	24	23	25	26	26	25
	C	23	20	40		17	28	26	32	30	30	28	31	28	29
	U	24	27	18		14	22	21	23	22	25	23	21	23	24
		a	a	c	STOP	g	g/c	g	c	c	c	c	c	c	c
Vertebrate	A	33	32	24		33	26	35	30	33	35	31	35	30	30
viruses	G	27	13	21		26	24	21	23	21	20	19	21	22	22
n = 267	C	16	20	24		15	21	21	22	26	18	23	16	23	22
	U	24	35	31		27	29	23	25	20	27	28	28	23	27
		a	u	u	STOP	a	u	a	a	a	a	a	a	a	a
Drosophila	A	24	44	17		40	29	28	24	27	31	27	38	29	30
n = 181	G	27	12	29		35	22	23	24	24	22	22	22	19	17
	C	24	15	31		12	26	22	24	27	22	23	22	25	28
	U	25	29	23		13	23	27	28	23	25	28	19	26	24
		g	a	c	STOP	A/G	a	a	u	a/c	a	u	a	a	a
Other	A	25	38	25		52	25	30	35	36	33	29	28	32	30
invertebrates	G	26	15	18		18	12	21	23	17	13	13	21	16	19
n = 126	C	22	21	27		7	25	22	21	20	19	26	20	24	17
	U	27	26	30		22	37	26	21	28	32	33	31	29	33
		u	a	u	STOP	A	u	a	a	a	a	u	u	a	u
Monocot plants	A	22	40	17		37	16	29	35	19	17	28	34	17	13
n = 86	G	42	12	21		31	34	27	26	27	29	19	20	22	33
	C	20	27	44		9	27	20	20	31	29	23	18	33	30
	U	16	22	17		22	23	24	19	23	25	30	28	28	23
		g	a	c	STOP	a	g	a	a	c	g/c	u	a	c	g
Dicot plants	A	33	40	13		39	26	26	28	31	33	29	35	32	29
n = 210	G	28	9	23		26	15	19	18	19	17	18	16	14	16
	C	15	22	30		7	22	16	17	21	15	17	18	21	19
	U	24	28	34		29	36	40	38	30	35	37	31	33	36
		a	a	u	STOP	a	u	u	u	a	u	u	a	u	u
Yeast	A	33	41	30		36	28	27	37	38	31	27	35	32	33
n = 417	G	25	12	16		27	20	19	19	19	15	15	20	20	17
	C	15	17	18		8	21	14	16	13	17	17	13	17	14
	U	27	29	35		29	31	30	28	30	38	41	32	32	36
		a	a	u	STOP	a	u	u	a	a	u	u	a	a/u	u
Dictyostelium	A	36	43	32		70	32	40	55	56	31	50	62	46	54
n = 53	G	17	6	2		4	2	4	0	6	6	2	12	2	
	C	13	13	13		4	17	11	11	8	6	10	6	15	13
	U	34	38	53		23	47	47	30	37	58	35	31	27	31
		a	A/U	U/A	STOP	A	A/U	A/U	A/U	A/U	A/U	A/U	A	a/u	A/U
Protozoa	A	25	39	20		48	26	41	36	38	27	29	38	40	33
n = 113	G	28	7	21		23	19	19	22	14	13	20	20	14	15
	C	19	13	33		9	21	18	12	15	19	13	14	20	19
	U	28	41	26		20	34	22	29	33	41	37	27	26	34
		g/u	U/A	c	STOP	a	u	a	a	a	u	u	a	a	u

preferred in all the other eukaryotic groups. UAG is the least frequently used termination codon for all groups except Drosophila. A survey of the relative GC content of the sequences flanking the termination codons suggested that the frequency of termination codons containing G (i.e. UAG and UGA) might be related to the GC content of the mRNA. To examine this possible relationship, the regression of the sum of UAG and UGA frequency (Table 5) on the average GC content of positions +1 to +10 was determined. The regression (0.88) is highly significant ($P < 0.01$) consistent with the hypothesis that termination codon usage is highly influenced by GC content. Brown and coworkers (21) previously noted this tendency in a smaller dataset. As was seen for the 5′ untranslated region, the 3′ untranslated region of vertebrate virus mRNAs are more AU rich than cellular vertebrate mRNAs. Consistent with the general relationship of GC content and termination codon usage, Vertebrate virus coding sequences are more frequently terminated by UAA than cellular vertebrate mRNAs. Differences in GC content does not explain the large and variable differences between UAG and UGA usage (Table 5).

A previous compilation of eukaryotic termination sequences indicated a purine bias at the +1 position immediately downstream of the termination codon (21). Our compilation of the sequences flanking eukaryotic stop codons revealed a more complex situation (Table 6). Although all taxa exhibit a purine bias, most of the groups also exhibit an appreciable frequency of U (>20%) at the +1 position. Overall, eukaryotes exhibit a low frequency of C (< 17%) at this position. Since these biases may differ between the three termination codons, the nucleotide frequencies for each termination codon class was examined separately (Table 7). All non-vertebrates eukaryotes were combined for this analysis to provide an adequate sample size. Each of the three termination codon classes for vertebrates and other eukaryotes have a preference for purines (ranging from 63% to 71%) at the +1 position with A being more frequent than G except for the UGA-vertebrate class. The frequency of C is particular low at this position for the UAA-Other Eukaryotes class and exhibits a higher frequency at the +2 position than the +1 position for all classes. The frequency of U is lower at the +1 position compared with the +2 position for all classes except for the UGA-Other Eukaryotes class. Differences in the GC content of +1 to +10 are observed between codon classes within each of the two taxonomic groups. The UAA mRNAs are more AU rich than the UAG and UGA classes analogous to the association of termination codon usage and GC content across taxa noted above.

Although strong frequency biases are also observed in the first three nucleotide positions upstream of the stop codon (Table 6 and 8), analysis of other adjacent codon positions indicate that these biases are not unique to the last amino acid codon (see below).

Table 7 Nucleotides frequencies flanking the three stop codons in vertebrates and non-vertebrates

Vertebrates

	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
A	31	38	19	0	100	100	39	25	27	25	29	21	27	23	24	24
G	24	17	23	0	0	0	29	25	24	24	21	21	22	26	25	24
C	21	17	36	0	0	0	15	26	27	27	27	28	24	27	25	26
U	24	28	23	100	0	0	16	24	23	23	23	29	26	24	27	26
	a	a	c	U	A	A	a	c	a/c	c	a	u	a	c	u	c/u
A	25	37	15	0	0	100	29	22	28	15	22	22	22	20	22	21
G	28	17	29	0	100	0	42	32	26	27	25	25	26	26	25	27
C	25	22	45	0	0	0	15	27	27	36	31	31	30	35	32	31
U	22	24	12	100	0	0	13	19	20	22	23	23	22	19	21	21
	g	a	c	U	G	A	g	g	a	c	c	c	c	c	c	c
A	28	32	14	0	100	0	34	21	20	16	21	24	21	20	24	21
G	24	14	30	0	0	100	29	24	35	26	28	24	28	27	28	26
C	22	22	36	0	0	0	23	32	24	32	32	30	29	30	27	26
U	26	31	21	100	0	0	14	23	21	25	19	23	23	23	21	24
	a	a	c	U	A	G	a	c	g	c	c	c	c	c	g	g/c

Other Eukaryotes

	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
A	27	39	22	0	100	100	39	27	33	35	36	29	29	34	35	31
G	29	11	17	0	0	0	30	17	18	20	18	16	16	20	17	17
C	16	17	29	0	0	0	6	23	16	18	16	17	17	17	19	19
U	28	32	32	100	0	0	24	33	33	28	30	38	39	29	29	33
	g	a	u	U	A	A	a	u	a/u	a	a	u	u	a	a	u
A	34	43	23	0	0	100	37	24	34	34	35	30	27	34	28	30
G	25	10	23	0	100	0	26	26	23	18	22	18	18	21	20	17
C	19	22	23	0	0	0	10	25	16	17	18	22	21	18	22	18
U	23	26	30	100	0	0	27	26	26	31	26	31	34	27	30	34
	a	a	u	U	G	A	a	g/u	a	a	a	u	u	a	u	u
A	29	45	21	0	100	0	47	29	28	26	25	31	26	35	26	27
G	27	13	28	0	0	100	19	16	24	29	20	18	18	18	17	22
C	21	16	26	0	0	0	13	21	21	20	27	18	25	17	25	22
U	24	26	25	100	0	0	21	33	26	26	28	32	32	30	32	29
	a	a	g	U	A	G	a	u	a	g	u	u	u	a	u	u

Multipositional analysis of termination sequences

Triplet nucleotide frequencies immediately upstream of the termination site may be influenced by a multitude of factors including: codon bias, amino acid composition, protein structure, normal biases encountered in the genome, and selection for sequences compatible with translation termination. The triplet AAG, encoding lysine, is the most abundant triplet immediately upstream of the termination site for vertebrate and non-vertebrate eukaryotes. Lysine is the most abundant amino acid at the carboxyl terminus of both eukaryotic groups. However this is not unique to this position; an equally high frequency of lysine codons are observed at the second and third to the last codons before the termination site as well.

Analysis of doublet frequencies at position −5 and −4 indicated a paucity of GG and GA for both major eukaryotic groups: 4.1% GG in vertebrates and 7.5% GG in non-vertebrates. Both frequencies are approximately half of the average observed GG + GA frequencies for the second and third positions of the 4th, 3rd, and 2nd to the last codons (data not shown).

The frequency bias at the +4 position (Tables 6 & 7 and reference 22) may be the result of selection for nucleotides which optimize (or regulate) termination of translation. Alternatively, this bias may simply be a reflection of natural quadruplet biases which include termination codons. In order to examine the latter hypothesis, quadruplet frequencies were computed in the region from +4 to +13 and then compared to the quadruplet frequencies containing the functional stop codons. A consistent trend is observed for the 3′ nucleotide adjacent to the stop codons (Table 8). The frequencies of the two purines at this position are greater adjacent to functional stop codons than adjacent to downstream stop codons and the frequencies of the two pyrimidines are always less frequent adjacent to functional stop codons than adjacent to downstream stop codons. This is true for all termination codons and the two eukaryote groups. For both eukaryote groups the UAA mRNAs shows the strongest purine bias at this position.

Table 8 Frequency at 3′ position downstream of functional stop codons (+) compared downstream stop codons (-)

	A		G		C		U	
	+	-	+	-	+	-	+	-
Vertebrates								
UAA	39 > 29		29 > 26		15 < 19		16 < 25	
UGA	29 > 25		42 > 31		15 < 27		23 < 30	
UAG	34 > 30		29 > 22		23 < 30		14 < 17	
Non-vertebrates								
UAA	39 > 34		30 > 16		6 < 15		24 < 35	
UGA	37 > 35		26 > 19		10 < 15		27 < 30	
UAG	47 > 37		19 > 14		13 < 21		21 < 28	

Analysis of doublet, triplet, and quadruplet frequencies downstream of the termination site were calculated and compared with expected frequencies. The expected frequencies were calculated using method-II. The correction factors were calculated from +10 to +19 of the 3′ untranslated region. Although numerous observed frequencies deviate significantly from expected frequencies (data not shown), almost all of the deviations can be explained by the heterogeneous mixture of GC biased verses AU biased mRNAs. Perhaps more importantly no triplet or quadruplet combination is dominant in the region immediately downstream of the termination site. The most frequent triplet occurred in only 5.2% of all mRNAs and the most frequent quadruplet in 2.1% of all mRNAs. (These frequencies are 4−5 fold lower than the most frequent triplet and quadruplet upstream of the start site.) Therefore, there are no dominant sequence motifs immediately downstream of the termination site. On the other hand there are a few triplet sequences immediately 3′ adjacent to the termination site that are rare. Only six vertebrate mRNAs contain either CGA and CGC at this position; 19 are expected. CGG and CGU triplets are very rare (5 in total) in non-vertebrates.

DISCUSSION

Diverse sequences support translation initiation

Analysis of the sequence flanking translation initiation sites indicates a large degree of variation both within and between the major eukaryotic groups. Yet embedded within this diversity are a number of general conserved features including: a strong preferences for purines at the −3 position, a periodical increase in the frequency of G at positions −9, −6, and −3, and a preference for A or C at positions −5, −4, −2, and −1. These similarities argue in favor of a conserved mechanism of initiation site selection. However, the diversity of sequence combinations surrounding translational initiation suggests that most sequence context are functionally acceptable in an appropriate cellular context. Indeed, many bonafide translation initiation sites fit rather poorly to the consensus sequence. This of course does not mean sequence contexts are functionally equivalent. Extrapolating from mutagenesis experiments of sequence surrounding start codons of the rat preproinsulin gene and the Drosophila alcohol dehydrogenase gene, we speculate that many functional eukaryotic translation initiation start sites may support a relative low level of translation initiation. Germane to this is the curious decline over the past decade in the estimated frequency of A at the −3 position upstream of the start codon that has occurred

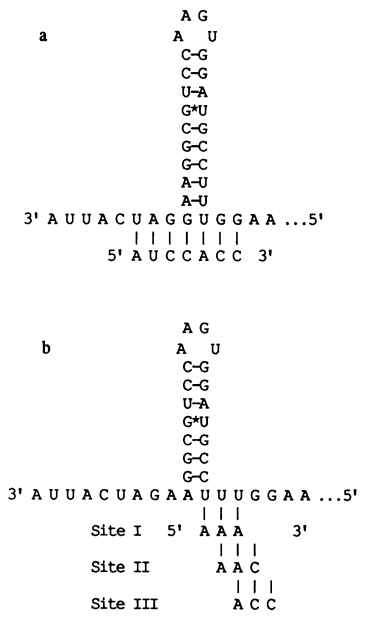

Fig. 1. Model for the interaction of eukaryotic mRNAs with the ribosomal 18S RNA. (A) Model proposed by Sargan and coworkers (28). (B) Model presented herein. Binding site I and II in Model B requires a shorter stem (eliminating one or two AU base pairs in the stem shown in Model A). The archetypal mRNA sequence which may bind to each of the three binding sites is shown below the 18S rRNA sequence. Allowing GU wobble base pairs expands the number of mRNA sequences which can bind to sites I, II, and III to a total of twenty-four triplets.

in published compilations of these sequences. We speculate that the first mRNA sequences compiled were biased towards highly expressed genes which were natural targets for cDNA cloning due to their relative abundant mRNAs. To a large extent the avalanche of recently sequenced mRNAs has been a consequence of more sophisticated methods to isolate cDNA clones of mRNAs expressed at low levels. Assuming that transcription and translation rates are generally correlated, the significantly higher frequency of A at the −3 position observed in Kozak's initial survey could be due to a bias towards mRNAs that are translated at relatively high rates.

Consensus sequence versus optimal sequence

Kozak originally suggested CCA/GCCAUGG (1, 2) as the consensus sequence without defining criteria for determining consensus. Later, Kozak extended this sequence to GCCGCC-A/GCCAUGG (8). It is interesting to note the number of vertebrate mRNA exhibiting either one of the proposed consensus sequences is very low. Of the 2,595 mRNA sequences only 76 (3%) exhibit the sequence CCA/GCCAUGG and 6 (0.2%) exhibit the sequence GCCGCCA/GCCAUGG. In fact what Kozak has convincingly demonstrated is that the later sequence provides exceptionally high levels of translation (8). We argue that the concepts of consensus and optimal are quite different and should not be equated with each other. *Consensus* is strictly a statistical term whereas *optimal* is a functional term. Moreover, the term optimal is somewhat problematic because it is often incorrectly equated with 'high' or 'maximal'. Optimal really means *best* which for a particular mRNA and protein could mean a *low* level of translation. These distinctions are particularly important for translation initiation sites because of the common practice of using published consensus sequences as a standard for determining the likelihood that a methionine codon is a start codon.

A revised model of the interaction of the 18S rRNA and mRNA sequences at the initiation site

Presently, the interaction of the eukaryotic preinitiation complex with the mRNA initiation site is poorly understood. In bacteria the Shine-Dalgarno (SD) mRNA sequence base-pairs with a 5−6 base region (anti-SD) very near the 3′ end of the 16S rRNA (24). This interaction strongly influences initiation site selection (25, 26). The SD sequence is typically located between −15 and −4 of the mRNA. The Interbas program described herein easily detects the Shine-Dalgarno sequences in bacterial mRNAs (data not shown). However, we are certain from our analyses that no analogous sequence motif is found in the same region of eukaryotic mRNAs. Although the bacterial 16S rRNA shares a very high degree of sequence identity with the corresponding eukaryotic 18S rRNA, their 3′ terminal sequences contain two differences (27). First, eukaryotes exhibit a precise deletion of the anti-SD sequence (CCUCC). Secondly, bacteria contain an AA doublet which base-pairs in a secondary structure with a downstream UU doublet; the order of AA and UU doublets in the sequence is reversed in eukaryotes (Fig. 1a). Sargan and coworkers (28) proposed that the base of the 3′ terminal 18S rRNA stem loop structure base-pairs with the −5 to −1 region of eukaryotic mRNAs (Fig. 1a). This model was based upon the notion that eukaryotic mRNAs closely resembled their originally reported consensus sequence, AUCCACC (−7 to −1 from the start codon). However, as detailed herein, this is not a general consensus sequence for eukaryotes.

Models which describe the interaction of eukaryotic mRNA and the ribosome must take in to account the diverse combination of sequences allowed in the mRNA sequences immediately flanking the start codon. In particular, it should be compatible with the most frequent −321 triplet (AAA) in non vertebrates and vertebrate viruses. A search of the terminal 3′ stem-loop structure and immediate flanking sequence of the 18S rRNA revealed a single UUU sequence. By analogy with the structure of the 16S rRNA sequence of *E. coli*, this sequence would occur at the end of the terminal stem (Fig 1a). This sequence is particularly interesting because two of the Us are the UU doublet mentioned above which resides on the opposite side of this stem in bacterial 16S rRNA. Moreover this triplet is adjacent to the GG doublet proposed by Sargan and coworkers to base pair with mRNA (28). We propose that the −321 of mRNA may base-pair with one of three triplet frames found in a five nucleotide segment (3′-UUUGG-5′) of the 18S rRNA (Fig 1b). Site-I would perfectly base pair with AAA, the most frequent −321 triplet in non-vertebrates and site-III would base-pair with ACC, the most frequent −321 triplet in vertebrates. Site-III is part of the Sargan's proposed sequence and may span the terminal stem-loop structure of the 18S rRNA as originally proposed (28). Site-II would bind to the sequence AAC. Although Watson-Crick base-pairing would only account for binding one mRNA −321 triplet at each of the three 18S rRNA binding sites, allowing one or more GU 'wobble' base-pairs expands the number of triplet base-pairs to twenty-four. The twenty-four −321 triplets account for 81% of all vertebrate and 67% of all non-vertebrate mRNA sequences. A potential problem of this model is that binding to site-I or II would require that the one or both of the Us are not base-paired in the the stem-loop structure. This constitutes a small

loss (approximately −1.2 kcal/mole) to the strength of the stem-loop structure (29). Binding to site-III could occur with the UU doublet base-paired in the stem loop structure. Thus, we envision that the 18S rRNA stem-loop structure may exist in equilibrium with the three alternative forms as consequence of 'breathing' at the end of the stem. This equilibrium may be further influenced by base-pair competition with the −321 triplet of a particular mRNA. An implication of this model is that the three alternative 18S rRNA binding sites may be an adaptation to cope with the diversity of AU biased versus GC biased mRNAs which occurs both within and between species of organisms. It should be noted our triplet model is compatible with Kozak's suggestion that GCC triplets at −987 and −654 may act to phase the scanning of the 40S ribosomal subunit to the proper reading frame (23). We suggest that our three proposed binding sites in the 18S rRNA may act as the phasing element since the most frequent −654 triplets can base-pair with one of the three sites.

Interaction of the ribosome with the mRNA termination sequences

Prokaryotes and eukaryotes both exhibit a strong bias immediately 3′ to stop codons (21, 22 and data presented herein). However, prokaryotes strongly prefer U whereas eukaryotes prefer purines or purines and U. Both groups share a strong bias against C at this position. Shine and Dalgarno (24) noted several years ago that the last six nucleotides at the 3′ end of the eukaryotic 18S rRNA (3′ $_{HO}$AUU ACU−5′) contained two triplets which could potentially base-pair with the three termination codons. The presence of A or G at the 3′ position immediately downstream of the stop codon does not add to the stability of this proposed interaction since the anticodon triplets in the 18S rRNA both contain A at the complementary positions (i.e. 3′-AUUA-5′ and 3′-ACUA-5′). The mechanism of termination in prokaryotes is still not clearly understood. Although the 3′ end of the 16S rRNA has been implicated in translational termination in *E. coli* (24) other sites (not conserved in the 18S rRNA) which may base-pair with the termination codons have recently been shown to be necessary for termination (30). The mechanism of termination in prokaryotes and eukaryotes may be quite different, inasmuch the termination release factors of *E. coli* and mammals are apparently not homologous (31).

Although strong biases for doublet, triplet, and quadruplet nucleotide frequencies were not observed downstream of the termination codons, we did note a paucity of the CG dinucleotide immediately downstream of the termination codon in all eukaryotes. Part of the bias against CG dinucleotides at this position in vertebrates can be explained by the general bias against CG. However, other taxa which do not show a general bias against CG nonetheless, exhibit a strong bias against CG downstream of termination codons. In particular, no yeast mRNA has been reported to contain either CGG or CGU immediately downstream of a stop codon terminating a major protein coding region. However, we have noted that CGG is found immediately 3′ adjacent to the URF 4 (upstream open reading frame) termination codon in the GCN4 leader sequence. URF4 is critically involved in the translation repression of the GCN4 protein (32,33). URF4 presents a very strong block to reinitiation downstream at the major GCN4 start codon. Miller and Hinnebusch (33) have shown that the sequences immediately downstream of the URF4 termination codon are largely responsible for the unique characteristics of URF4 compared with other URFs. The nature of URF4 block of downstream reinitiation at the major GCN4 start codon is unknown but may involve aberrant termination due to the presence of the CGG triplet downstream of the termination codon.

ACKNOWLEDGEMENT

This work was supported by a grant from a Public Health Service grant, GM34170.

REFERENCES

1. Kozak, M. (1981) Nucl. Acids Res. **9,** 5233−5252.
2. Kozak, M. (1984) Nucl. Acids Res. **12,** 857.
3. Cavener, D. R. (1987) Nucl. Acids Res. **15,** 1353−1361.
4. Laz, T., Clements, J., and Sherman, F. (1987) Translational regulation of gene expression, J. Ilan, eds. (New York: Plenum Publishing Corp.) pp. 413−429.
5. Hamilton, R., Watanabe, C. K., and de Boer, H. A. (1987) Nucl. Acids. Res. **15,** 3581−3593.
6. Lutcke, H., Chow, K., Mickle, F., Moss, K., Kern, H., and Scheele, G. (1987) EMBO J. **6,** 43−48.
7. Kozak, M. (1986) Cell **44,** 283−292.
8. Kozak, M. (1987) J. Mol. Biol. **196,** 947−950.
9. Baim, S. B., and Sherman, F. (1988) Mol. Cell. Biol. **8,** 1591−1601.
10. Cigan, A. M., Pabich, E. K., and Donahue, T. F. (1988) Mol. Cell. Biol. **8,** 2964−2975.
11. Kozak, M. (1989) Mol. Cell. Biol **9,** 5073−5080.
12. Feng, Y., Gunter, L. E., Organ, E. L., and Cavener, D. R. (1991) Mol. Cell. Biol. **11,** 2149−2153.
13. Valle, R. P., and Morch, M. D. (1988) FEBS Lett. **235,** 1−15.
14. Craigen, W. J., and Caskey, C. T. (1987) Cell **50,** 1−2.
15. Brierley, I., Digard, P., and Inglis, S. C. (1989) Cell **57,** 537−547.
16. Peabody, D. S., and Berg, P. (1986) Mol. Cell. Biol. **6,** 2695−2703.
17. Peabody, D. S., Subramani, S., and Berg, P. (1986) Mol. Cell. Biol. **6,** 2704−2711.
18. Werner, M., Feller, A., Messenguy, F., and Pierard, A. (1987) Cell **49,** 805.
19. Mueller, P., and Hinnebusch, A. (1986) Cell **45,** 201−207.
20. Kozak, M. (1987) Mol. Cell. Biol. **7,** 3438.
21. Brown, C. M., Stockwell, P. A., Trotman, C. N. A, and Tate, W. P. (1990) Nucl. Acids Res. **18,** 6339−6345.
22. Brown, C. M., Stockwell, P. A., Trotman, C. N. A, and Tate, W. P. (1990) Nucl. Acids Res. **18,** 2079−2086.
23. Kozak, M. (1987) Nucl. Acids Res. **15,** 8125−8148.
24. Shine, J., and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA **71,** 1342−1346.
25. Gold, L. (1988) Ann. Rev. Biochem. **57,** 199−233.
26. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S., and Stormo, G. (1981) Ann. Rev. Microbiol. **35,** 365−403.
27. Van Charldorp, R., and Van Knippenberg, P. H. (1982) Nucl. Acids Res. **10,** 1149−1158.
28. Sargan, D. R., Gregory, S. P., and Butterworth, P. H. W (1982) FEBS Lett. **147,** 133−136.
29. Saenger, W. (1983) Principles of Nucleic Acid Structure. In Springer Advanced Texts in Chemistry, C. Cantor, eds. (New York: Springer-Verlag) pp. 146−152.
30. Murgola, E. J., Hijazi, K. A., Goringer, H. U., and Dahlberg, A. E. (1988) Proc. Natl. Acad. Sci. USA **85,** 4162−4165.
31. Lee, C. C., Craing, W. J., Muzny, D. M., Harlow, E., and Caskey, C. T. (1990) Proc. Natl. Acad. Sci. USA **87,** 3508−3512.
32. Williams, N. P., Mueller, P. P., and Hinnebusch, A. G. (1988) Mol. Cell. Biol. **8,** 3827−3836.
33. Miller, P. F., and Hinnebusch A G (1989) Genes Devel. **3,** 1217−1225.